Appl. Math. Lett. Vol. 3, No. 3, pp. 13–18, 1990
Printed in Great Britain.

0893-9659/90 $3.00 + 0.00

Application of Adjoint Operators to Neural Learning

J. BARHEN[1,2], N. TOOMARIAN[1] AND S. GULATI[1]

[1] Jet Propulsion Laboratory
[2] Division of Engnieering and Applied Sciences

(Received May 1990)

Abstract. A new methodology for neural learning of nonlinear mappings is presented. It exploits the concept of *adjoint operators* to enable a fast global computation of the network's response to perturbations in all system parameters.

1. INTRODUCTION

A considerable effort has recently been devoted to the development of efficient computational methodologies for learning. Attention has largely focussed on the back-propagation algorithm because of its simplicity, generality and the promise that it has shown in regard to various applications [9,12]. More recently, Pineda [8] has derived a generalization to back-propagation for recurrent networks. In a similar vein, Williams and Zipser [13] have presented algorithms for learning tasks with temporal dependencies. Pearlmutter [7] has proposed a similar technique which minimizes an error functional between output and targeted temporal trajectories. In a significantly different approach, Barhen, Gulati and Zak [3,4] recently introduced neural formalisms to efficiently learn nonlinear mappings using a new mathematical construct, i.e., terminal attractors [14]. Terminal attractor representations were used not only to ensure infinite local stability of the encoded information, but also to provide a qualitative as well as quantitative change in the nature of the learning process. In particular, they imply loss of Lipschitz conditions at energy function minima, which results in a dramatic increase in the speed of learning.

The development of learning algorithms is generally based upon the minimization of a "neuromorphic" energy-like function. A fundamental requirement of all previously mentioned methods is the computation of the gradient of this objective function with respect to the various parameters of the neural architecture, e.g., synaptic weights, neural gain, etc. In the present paper we introduce a new methodology for their efficient analytical computation, as a single solution of a set of "adjoint" equations. We have already successfully used adjoint operators in some of our earlier work in the fields of energy economy modeling [1] and nuclear reactor thermal hydraulics [2,11] at the Oak Ridge National Laboratory, where the concept flourished during the past decade [5,6].

The research described in this paper was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by agencies of the U.S. Department of Defense, and by the Office of Basic Energy Sciences of the U.S. Department of Energy, through agreements with NASA. We wish to acknowledge discussions with Michail Zak and Fernando Pineda.

2. ADJOINT OPERATORS

Consider, for the sake of generality, that a problem of interest is represented by the following system of N coupled nonlinear equations

$$\bar{\varphi}(\bar{u}, \bar{p}) \;=\; 0 \tag{2.1}$$

where $\bar{\varphi}$ denotes a nonlinear operator. If differential operators appear in Eq. (2.1), then a corresponding set of boundary and/or initial conditions to specify the domain of φ must also be provided. The learning model discussed in this paper focuses on the adiabatic approximation only (steady state networks). Nonadiabatic learning algorithms will be discussed in a forthcoming article [10].

Let $\bar{u}$ and $\bar{p}$ represent the N-vector of dependent variables and the M-vector of system parameters, respectively. We will assume that generally $M >> N$ and that elements of $\bar{p}$ are, in principle, independent. Furthermore, we will also assume that, for a specific choice of parameters, a unique solution of Eq. (2.1) exists. Hence, $\bar{u}$ is an implicit function of $\bar{p}$. A system response, R, represents any result of the calculations that is of interest. Specifically

$$R \;=\; R(\bar{u}, \bar{p}) \tag{2.2}$$

i.e., R is a known nonlinear function of $\bar{p}$ and $\bar{u}$ and may be calculated from (2.2) when the solution $\bar{u}$ in Eq. (2.1) has been obtained for a given $\bar{p}$. The problem of interest is to compute the "sensitivities" of R, i.e., the derivatives of R with respect to parameters p_μ, $\mu = 1, \cdots, M$. By definition

$$\frac{dR}{dp_\mu} \;=\; \frac{\partial R}{\partial p_\mu} + \frac{\partial R}{\partial \bar{u}} \cdot \frac{\partial \bar{u}}{\partial p_\mu} \tag{2.3}$$

Since the response R is known analytically, the computation of $\partial R/\partial p_\mu$ and $\partial R/\partial \bar{u}$ is straightforward. The quantity that needs to be determined is the vector $\partial \bar{u}/\partial p_\mu$. Differentiating the state equations (2.1), we obtain a set of equations to be referred to as "forward" sensitivity equations

$$\frac{\partial \bar{\varphi}}{\partial \bar{u}} \cdot \frac{\partial \bar{u}}{\partial p_\mu} \;=\; -\frac{\partial \bar{\varphi}}{\partial p_\mu} \tag{2.4}$$

To simplify the notations, we are omitting the "transposed" sign and denoting the N X N forward sensitivity matrix $\partial \bar{\varphi}/\partial \bar{u}$ by A, the N-vector $\partial \bar{u}/\partial p_\mu$ by ${}^\mu\bar{z}$ and the "source" N-vector $-\partial \bar{\varphi}/\partial p_\mu$ by ${}^\mu\bar{s}$. Thus

$$A\,{}^\mu\bar{z} \;=\; {}^\mu\bar{s} \tag{2.5}$$

Computation of the response gradient using the forward sensitivity equations would require solving a system of N nonlinear algebraic equations for each parameter p_μ, since the source term in Eq. (2.5) explicitly depends on μ. This difficulty is circumvented by introducing adjoint operators. Let A^* denote the formal adjoint of the operator A [1,11][1]. The adjoint sensitivity equations can then be expressed as

$$A^*\,{}^\mu\bar{z}^* \;=\; {}^\mu\bar{s}^*. \tag{2.6}$$

By definition, for algebraic operators[2]

$${}^\mu\bar{z}^* \cdot (A\,{}^\mu\bar{z}) \;=\; {}^\mu\bar{z}^* \cdot {}^\mu\bar{s} \;=\; {}^\mu\bar{z} \cdot (A^*\,{}^\mu\bar{z}^*) \;=\; {}^\mu\bar{z} \cdot {}^\mu\bar{s}^* \tag{2.7}$$

[1] Adjoint operators can only be considered for densely defined linear operators on Banach spaces. For the neural application under consideration we will limit ourselves to real Hilbert spaces. Such spaces are self-dual.

[2] The domain of an adjoint differential operator is determined by selecting appropriate adjoint boundary conditions. The associated bilinear form evaluated on the domain boundary must generally be also included.

Since Eq. (2.3) can be rewritten as

$$\frac{dR}{dp_\mu} = \frac{\partial R}{\partial p_\mu} + \frac{\partial R}{\partial \bar{u}} {}^{\mu}\bar{z} , \tag{2.8}$$

if we identify

$$\frac{\partial R}{\partial \bar{u}} \equiv {}^{\mu}\bar{s}^{*} \equiv \bar{s}^{*} \tag{2.9}$$

we observe that the source term of the adjoint equations is independent of the specific parameter p_μ. Hence, *the solution of a single set of adjoint equations will provide all the information required to compute the gradient of R with respect to all parameters.* To underscore that fact we shall denote ${}^{\mu}\bar{z}^{*}$ as $\bar{v}$. Thus

$$\frac{dR}{dp_\mu} = \frac{\partial R}{\partial p_\mu} + \bar{v} \cdot {}^{\mu}\bar{s}. \tag{2.10}$$

3. Applications to Neural Learning

We formalize a neural network as an adaptive dynamical system whose temporal evolution is governed by the following set of coupled nonlinear differential equations

$$\dot{u}_n + \kappa_n u_n = \sum_m T_{nm}\, g(\gamma_m u_m) + {}^{k}I_n \tag{3.1}$$

where u_n represents the mean soma potential of the nth neuron and T_{nm} denotes the synaptic coupling from the $m-$th to the $n-$th neuron [3]. The constant κ_n characterizes the decay of neuron activity. The sigmoidal function $g(\cdot)$ modulates the neural response, with gain given by γ_m; typically, $g(\gamma z) = \tanh(\gamma z)$. The "source" term, ${}^{k}I_n$ encodes component contributions by the presented attractors ${}^{k}\bar{a}$ of the k-th training pattern via the expression

$${}^{k}I_n = \begin{cases} [{}^{k}a_n - g(\gamma_n u_n)]^{\beta} & \text{if } n \in S_X \\ 0 & \text{if } n \in S_H \cup S_Y \end{cases} . \tag{3.2}$$

The topographic input, output and hidden network partitions S_X, S_Y and S_H are architectural requirements related to the encoding of mapping-type problems. Details are given in [3]. In previous articles [3,4,14] we have demonstrated that in general, for $\beta = (2i+1)^{-1}$ and i a positive integer, such attractors have infinite local stability and provide opportunity for learning in real-time.

To proceed formally with the development of a learning algorithm, we consider an approach based upon the minimization of a constrained "neuromorphic" energy-like function E given by the following expression:

$$E(\bar{u}, \bar{\lambda}, \bar{p}) = \frac{1}{2} \sum_n \sum_m \omega_{nm} \left(T_{nm}^2 - T_{nm}\, T_{mn} \right) + \frac{1}{\alpha} \sum_k \sum_n {}^{k}\lambda_n \, {}^{k}\Gamma_n^{\alpha} \tag{3.3}$$

where the constraints are of the form

$${}^{k}\Gamma_n = \begin{cases} {}^{k}a_n - g(\gamma_n \, {}^{k}\tilde{u}_n) & \text{if } n \in S_X \cup S_Y \\ 0 & \text{if } n \in S_H \end{cases} \tag{3.4}$$

Typically, a positive value such as 2 is used for α. The weighting factor ω_{nm} is constructed in such a fashion, as to favor locality of computation. The indices n, m span over all neurons in the network. Lagrange multipliers corresponding to the $nk-$th constraint are denoted by ${}^{k}\lambda_n$. The superscript $\sim$ denotes quantities evaluated at steady state. The proposed

objective function includes contributions from two sources. First, it enforces convergence of every neuron in S_X and S_Y to attractor coordinates corresponding to the components in the input-output training patterns, thereby prompting the network to learn the underlying invariances. Secondly, it regulates the topology of the network by enforcing symmetry, and by minimizing interconnection strengths between distant synaptic elements to favor locality of computation.

Lyapunov stability requires an energy-like function to be monotonically decreasing in time. In our model the internal dynamical parameters of interest are the synaptic strengths T_{ij} of the interconnection topology, the characteristic decay constants κ_i, the gain parameters γ_i and the Lagrange multipliers ${}^k\lambda_i$. This implies that we require

$$\dot{E} = \sum_i \sum_j \frac{dE}{dT_{ij}} \dot{T}_{ij} + \sum_i \frac{dE}{d\kappa_i} \dot{\kappa}_i + \sum_i \frac{dE}{d\gamma_i} \dot{\gamma}_i + \sum_l \sum_i \frac{dE}{d^l\lambda_i} {}^l\dot{\lambda}_i < 0 \quad (3.5)$$

One can always choose, with $\tau_T > 0$

$$\dot{T}_{ij} = -\tau_T \frac{dE}{dT_{ij}} \quad (3.6)$$

where τ_T introduces an adaptive parameter for learning (see, e.g., [3,4]). Similar expressions can be constructed for $\dot{\kappa}$ and $\dot{\gamma}$, e.g.,

$$\dot{\kappa}_i = -\tau_\kappa \frac{dE}{d\kappa_i} \qquad \text{and} \qquad \dot{\gamma}_i = -\tau_\gamma \frac{dE}{d\gamma_i} \quad (3.7)$$

with τ_κ, $\tau_\gamma > 0$. Then, substituting in Eq. (3.5) and denoting tensor contraction by $\oplus$, one obtains

$$\nabla_\lambda E \oplus \dot{\lambda} < \tau_T (\nabla_T E \oplus \nabla_T E) + \tau_\kappa (\nabla_\kappa E \ominus \nabla_\kappa E) + \tau_\gamma (\nabla_\gamma E \oplus \nabla_\gamma E) \quad (3.8)$$

Without loss of generality, one can assume $\tau = \tau_T = \tau_\kappa = \tau_\gamma$.

The equations of motion for the Lagrange multipliers ${}^l\lambda_i$ must now be constructed in such a way that Eqn. (3.8) is strictly satisfied. In addition, when the constraints are satisfied, i.e., as ${}^l\Gamma_n \to 0$ in Eq. (3.4), we require that ${}^l\dot{\lambda}_i \to 0 \ \forall \ l$. We have adopted the following analytical model for the evolution of λ_i,

$${}^l\dot{\lambda}_i = \tau \frac{\Pi}{\Lambda + (1/(\Lambda + \theta))} {}^l[\nabla_\lambda E]_i \quad (3.9)$$

where $\Pi = \nabla_T E \oplus \nabla_T E + \nabla_\kappa E \oplus \nabla_\kappa E + \nabla_\gamma E \oplus \nabla_\gamma E$, $\Lambda = \nabla_\lambda E \oplus \nabla_\lambda E$ and θ is an arbitrary positive constant. It is straightforward to prove that this model fulfills the above requirements.

In relating adjoint theory to the neural learning algorithms, we identify the neuromorphic energy-like function, E in Eq. (3.3), with the system response. Let $\bar{p}$ denote the following system parameters:

$$\bar{p} = \{ T_{11}, \cdots T_{NN} \mid \kappa_1, \cdots \kappa_N \mid \gamma_1, \cdots \gamma_N \mid \cdots \} \quad (3.10)$$

The adiabatic solution to the nonlinear equations of motion (3.1), for each training pattern k, $k = 1, \cdots, K$ is given by

$${}^k\varphi_n({}^k\tilde{u}, \bar{p}) = -\kappa_n {}^k\tilde{u}_n + \sum_m T_{nm} g(\gamma_m {}^k\tilde{u}_m) + {}^kI_n = 0. \quad (3.11)$$

So, in principle, ${}^k\tilde{u}_n = {}^k\tilde{u}_n\,[T,\ \bar{\kappa},\ \bar{\gamma},\ {}^k a_n, \cdots]$. Using Eqs. (3.11), the forward sensitivity matrix can be computed and compactly expressed as

$$\begin{aligned} {}^kA_{nm} &= \frac{\partial\, {}^k\varphi_n}{\partial\, {}^k\tilde{u}_m} = [\,-\kappa_n + \frac{\partial\, {}^kI_n}{\partial\, {}^k\tilde{u}_m}]\,\delta_{nm} + T_{nm}\, {}^k\hat{g}_m\, \gamma_m \\ &= {}^k\eta_n\,\delta_{nm} + \gamma_m\, {}^k\hat{g}_m\, T_{nm} \end{aligned} \tag{3.12}$$

where $\hat{g}_m$ represents the derivative of g_m with respect to u_m. The adjoint sensitivity matrix is

$${}^kA^*_{nm} = {}^k\eta_m\delta_{mn} + \gamma_n\, {}^k\hat{g}_n\, T_{mn}. \tag{3.13}$$

Using Eqs. (2.9) and (3.3), we can compute the adjoint source:

$${}^ks^*_n = -\,{}^k\lambda_n\, {}^k\Gamma_n^{\alpha-1}\, \gamma_n\, {}^k\hat{g}_n \tag{3.14}$$

The system of adjoint equations can then be constructed using Eqs. (3.13) and (3.14), to yield :

$$\sum_m [\ {}^k\eta_m\,\delta_{mn} + T_{mn}\, {}^k\hat{g}_n\,\gamma_n\ ]\ {}^k\tilde{v}_m = -\,{}^k\lambda_n\, {}^k\Gamma_n^{\alpha-1}\,\gamma_n\, {}^k\hat{g}_n \tag{3.15}$$

Notice that the above system, (3.15), is linear in ${}^k\tilde{v}$. Furthermore, its components can be obtained as the equilibrium points, (i.e., $\dot{v}_i \rightarrow 0$) of the concomitant dynamical system

$$\dot{v}_n - {}^k\eta_n\, v_n = \gamma_n\, {}^k\hat{g}_n\,[\,\sum_m T_{mn}\, v_m + {}^k\lambda_n\, {}^k\Gamma_n^{\alpha-1}\,] \tag{3.16}$$

To proceed with our derivation of learning algorithms, we differentiate the steady-state equations (3.11) with respect to each parameter, p_μ, to obtain the forward source term, ${}^\mu s^k_n$:

$${}^\mu s^k_n = -\Big(\,[\,-{}^k\tilde{u}_i\,]\,\delta_{p_\mu,\kappa_i} + [\,\delta_{ni}\, g(\gamma_j\, {}^k\tilde{u}_j)\,]\,\delta_{p_\mu,T_{ij}} + [\,T_{ni}\, {}^k\hat{g}_i\, {}^k\tilde{u}_i + \frac{\partial\, {}^kI_n}{\partial\gamma_i}\,]\,\delta_{p_\mu,\gamma_i}\Big) \tag{3.17}$$

Substituting Eq. (3.17) in (2.10), and recalling that our abstract response corresponds here to the energy function E, yields

$$\frac{dE}{dp_\mu} = \frac{\partial E}{\partial p_\mu} + \sum_k {}^k\tilde{v}\,.^\mu\,\bar{s}^k \tag{3.18}$$

The explicit energy gradient contributions for parameters $p_\mu = T,\ \bar{\kappa},\ \bar{\gamma}$ immediately result :

$$\frac{dE}{dT_{ij}} = \omega_{ij}\, T_{ij} - \omega_{ji}\, T_{ji} - \sum_k {}^k\tilde{v}_i\, g(\gamma_j\, {}^k\tilde{u}_j) \tag{3.19}$$

$$\frac{dE}{d\kappa_i} = \sum_k {}^k\tilde{u}_i \sum_n {}^k\tilde{v}_n \tag{3.20}$$

$$\frac{dE}{d\gamma_i} = -\sum_k {}^k\lambda_i\, {}^k\Gamma_i^{\alpha-1}\, {}^k\hat{g}_i\, {}^k\tilde{u}_i + \sum_k\sum_n [\,T_{ni}\, {}^k\hat{g}_i\, {}^k\tilde{u}_i + \frac{\partial\, {}^kI_n}{\partial\gamma_i}\,]\, {}^k\tilde{v}_n \tag{3.21}$$

Substituting Eqs. (3.19)-(3.21) into Eqs. (3.6) and (3.7), we then obtain the complete learning dynamics.

4. CONCLUSIONS

In this paper we have presented a powerful theoretical framework for learning continuous nonlinear mappings using artificial neural networks. Central to our approach is the concept of *adjoint operators* which enables a fast computation of energy function gradients with respect to all system paramters using a single solution of the adjoint equations.

REFERENCES

1. R.G. Alsmiller, J. Barhen and J. Horwedel, *Energy* and references therein 9 (3), 239–253 (1984).
2. J. Barhen, D.G. Cacuci and J.J. Wagschal, *Nucl. Sci. Eng.* **81**, 23–44 (1982).
3. J. Barhen, S. Gulati and M. Zak, *IEEE Computer* **22** (6), 67–76 (1989).
4. J. Barhen, M. Zak and S. Gulati, *Proc. Neuro-Nimes*, 55–68 (1989).
5. D.G. Cacuci, C.F. Weber, E.M. Oblow and J.H. Marable, *Nucl. Sci. Eng.* **75**, 88–110 (1980).
6. E.M. Oblow, *ORNL TM 5815*, Oak Ridge National Laboratory, (1977).
7. B.A. Pearlmutter, *Neural Computation* **1** (2), 263–269 (1989).
8. F.J. Pineda, *Journal of Complexity* **4**, 216–245 (1988).
9. D.E. Rumelhart and J.L. McClelland, *Parallel and Distributed Processing*, MIT Press, (1986).
10. N. Toomarian and J. Barhen, Non-Adiabatic Learning in Neural Networks, *in Preparation.*
11. N. Toomarian, E. Wacholder and S. Kaizerman, *Nucl. Sci. Eng.* **99** (1), 53–81 (1987).
12. P. Werbos, Ph.D. Thesis, *Harvard Univ.* (1974).
13. R.J. Williams and D. Zipser, *Neural Computation* **1** (2), 270–280 (1989).
14. M. Zak, *Physics Letters A* **133**, 218–222 (1988).

California Institute of Technology, Pasadena, CA 91109